

A Structural Analysis of Human Granzyme B

Giulio Benedetti

7/3/2021

Contents

1. Abstract	1
2. Background	1
3. Methods	2
4. Results	2
5. Conclusions	8
6. References	9

1. Abstract

Background: Protein complexity arises from surprisingly simple building blocks, such as motifs and domains. Therefore, the functionality of topologically similar proteins can be predicted based on their elementary components. Such knowledge can also be transferred to newly discovered proteins, whose functions are not yet known. Due to the great variety of available Bioinformatics tools, it might be complicated to choose the right one for the addressed question. The goal of this report lies in demonstrating the proper use of some of these tools to conduct a structural analysis of the human Granzyme B.

Results: The human Granzyme B was identified as such via BLAST and its topological domains were predicted with multiple software. Its sequence was then three-dimensionally aligned to that of a similar protein, Granzyme H, which proved intermediate sequence conservation but high structure conservation. Next, an MSA was performed among various orthologs of Granzyme B and the corresponding phylogenetic tree was thereby reconstructed. This resulted in a visual interpretation of the evolutionary pattern of Granzyme B as well as an assessment of its conserved regions.

Conclusions: The immunological importance of Granzyme B as well as the Granzyme family is shown by the presence of these enzymes among several animals, where their sequences and binding sites might not necessarily stay unmodified. This variability could trigger enzyme paralogs as well as orthologs to target a large number of substrates, reflecting the proteomes of various pathogens. Finally, the Tripsin motif present in Granzyme B might cause the pyrolytic activity of this enzyme.

Keywords: protease Granzyme B, CD8+ T lymphocyte and NKT cell, BLAST, secondary structure analysis, multiple sequence alignment (MSA)

2. Background

The systematic analysis of protein structures has enabled researchers to connect the geometry of a protein with the functions it performs. As this knowledge broadens to newly discovered enzymes, their activity can be predicted based on their elementary components, such as motifs and domains, which often underpin the observed functionality. Similarly to incredible constructions built from simple LEGO bricks, several proteins arise from assembling the identical set of building blocks in different ways. A greater understanding of how proteins are assembled in terms of their topological elements could lead to important achievements, such as drug synthesis and plastic digestion.

Since the original research introduced traditionally accepted methods, such as the PAM¹ and BLOSUM² score matrices or the Chou-Fasman algorithm³, the techniques of choice have dramatically advanced to achieve a large variety of open source multifunctional as well as single-function software. In the same line, high-throughput sequencing, enzyme engineering through directed evolution and targeted mutagenesis, but also the simple fact that the scientific community is increasing in size, have produced a massive amount of data to be interpreted, demanding for a faster and more effective analytical framework. The amount of Bioinformatics tools available online might seem overwhelming, therefore young scientists must be prepared to ponder benefits and drawbacks of each of those tools and select the appropriate ones for the question to address.

Assuming such a mindset, in this report a structural analysis is conducted on an unknown protein, which is then identified as human Granzyme B. This protease is found in the cytosolic granules of CD8+ and TNK cells and is responsible for the apoptosis of target damaged cells in the context of the immune system.⁴⁵⁶

3. Methods

0.0.0.1 3.1. protein identification The mysterious protein ID6299 was obtained from the collection of proteins kindly provided by Dr. J. Weiner. It was identified through a BLAST search across the reference protein database (refseq_protein).⁷ The same query was also aligned through PSI-BLAST and within multiple databases (nr and swissprot).⁸ Additional information on the enzyme was collected from its UniProt overview.⁹

0.0.0.2 3.2. secondary structure prediction The online software PredictProtein was employed to predict the features of Granzyme B (secondary structure and sequence conservation).¹⁰ The predictions on aminoacid sequence and coiled-coil abundance were validated by two additional tools: Protscale¹¹ and COILS¹², respectively.

0.0.0.3 3.3. protein domains determination Further analyses on the structural domains of Granzyme B were conducted within the two large databases Pfam¹³ and PDB¹⁴. In the former, protein motifs and domains were characterised with a comparative approach, whereas in the latter it was visualised as a christallographic reconstruction.

0.0.0.4 3.4. secondary structure alignment The Protein Databank was blasted for similar proteins to Granzyme B¹⁵, from whose results human Granzyme H¹⁶ was selected. Next, the christallographic constructions of the two enzymes were found on the PDB database and their identifiers were passed to the DALI server so as to align their secondary structures with one another.¹⁷

0.0.0.5 3.5. multiple sequence alignment The NCBI BLAST tool was once again run across the nr database with the sequence of Granzyme B as a query. Its results were assessed in combination with the MSAs returned by PredictProtein and HomoloGene to select 7 elements to include in the MSA, which was generated with the EMBL-EBI tool clustalw¹⁸ and confirmed with clustal Omega¹⁹. Results of the latter were visualised with the local software Jalview.²⁰

0.0.0.6 3.6. sequence reconstruction the R packages *phangorn* 2.7.0²¹, *BiocManager* 1.30.16²² and *seqinr*²³ were used to create, optimise and bootstrap a phylogenetic tree out of the MSA. The resulting phylogenetic tree was compared with the dendrogram provided by clustalw upon multiple sequence alignment.²⁴²⁵

0.0.0.7 3.7. operational settings Unless reported, the operational settings of the aforementioned tools were kept at default.

4. Results

0.0.0.8 4.1. Granzyme B characterisation The unknown amino acid sequence was identified as Granzyme B isoform 1 preproprotein from *H. sapiens* (first hit). The same query produced analogous hits

when aligned with standard BLAST and PSI-BLAST across multiple databases (Figures 1 and 2). For further characterisation, the function and subcellular location of Granzyme B were studied: the protein is normally found whether in the extracellular matrix or within the cytosolic granules of CD8+ and NKT lymphocytes, which deliver it into the target cells to activate the apoptotic mechanism of caspase-independent pyroptosis (Figure 3).

granzyme B isoform 1 preprotein [Homo sapiens]

Sequence ID: [NP_004122.2](#) Length: 247 Number of Matches: 1

Range 1: 1 to 247 GenPept Graphics					▼ Next Match	▲ Previous Match
Score	Expect	Method	Identities	Positives	Gaps	
519 bits(1337)	0.0	Compositional matrix adjust.	247/247(100%)	247/247(100%)	0/247(0%)	
Query 1	MQPILLLLAFLLLPRADAGEIIGGHEAKPHSRPYMAYLMIWQDKSLKRCGGFLIRDDFVL			60		
Sbjct 1	MQPILLLLAFLLLPRADAGEIIGGHEAKPHSRPYMAYLMIWQDKSLKRCGGFLIRDDFVL			60		
Query 61	TAAHCWGSSINVTLGAHNIKEQEPTQQFIPVKRPPIPHPAYNPKNSNDIMLQLERKAKR			120		
Sbjct 61	TAAHCWGSSINVTLGAHNIKEQEPTQQFIPVKRPPIPHPAYNPKNSNDIMLQLERKAKR			120		
Query 121	TRAVQPLRLPSNKAQVPGQTCVAGWGQTAGPLGKHSHTLQEVKMTVQEDRKCESDLRHY			180		
Sbjct 121	TRAVQPLRLPSNKAQVPGQTCVAGWGQTAGPLGKHSHTLQEVKMTVQEDRKCESDLRHY			180		
Query 181	YDSTIELCVGDPEIKKTSFKGDSGGPLVCNKVAQGIVSYGRNNNGMPRACTKVSSFVHWI			240		
Sbjct 181	YDSTIELCVGDPEIKKTSFKGDSGGPLVCNKVAQGIVSYGRNNNGMPRACTKVSSFVHWI			240		
Query 241	KKTMKRY 247					
Sbjct 241	KKTMKRY 247					

Figure 1: Alignment of ID6299 with the first hit of the BLAST search. The score, E-value and percent identity suggest that the two sequences were matched flawlessly.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> granzyme B isoform 1 preprotein [Homo sapiens]	Homo sapiens	519	519	100%	0.0	100.00%	247	NP_004122.2
<input checked="" type="checkbox"/> granzyme B [Pan paniscus]	Pan paniscus	514	514	100%	0.0	98.38%	281	XP_003809164.2
<input checked="" type="checkbox"/> granzyme B isoform X1 [Pan troglodytes]	Pan troglodytes	513	513	100%	0.0	97.98%	281	XP_509879.2
<input checked="" type="checkbox"/> granzyme B isoform 2 [Homo sapiens]	Homo sapiens	484	484	93%	6e-173	99.57%	235	NP_001332940.1
<input checked="" type="checkbox"/> granzyme B isoform X2 [Pan troglodytes]	Pan troglodytes	478	478	93%	2e-169	97.39%	280	XP_016781433.1
<input checked="" type="checkbox"/> granzyme B isoform X1 [Gorilla gorilla gorilla]	Gorilla gorilla gorilla	473	473	100%	7e-168	94.33%	281	XP_004055081.1
<input checked="" type="checkbox"/> granzyme B isoform X1 [Pongo abelii]	Pongo abelii	469	469	100%	4e-166	93.52%	281	XP_002824669.1
<input checked="" type="checkbox"/> granzyme B isoform X2 [Gorilla gorilla gorilla]	Gorilla gorilla gorilla	458	458	93%	1e-161	93.48%	280	XP_018895534.1
<input checked="" type="checkbox"/> granzyme B isoform X1 [Nomascus leucogenys]	Nomascus leucogenys	457	457	100%	3e-161	91.20%	284	XP_030658664.1
<input checked="" type="checkbox"/> granzyme B isoform X2 [Pongo abelii]	Pongo abelii	454	454	93%	7e-161	92.61%	235	XP_024087456.1

Figure 2: First 10 hits of the BLAST search. All best matched organisms belong to the primates. However, the minuscule E-value of hit 1 and the appearance of two human isoforms (hit 1 and hit 4) indicates the human origin of the protein.

0.0.0.9 4.2. structural and topological prediction The results of the structural analysis showed that Granzyme B contains several β sheets and a small amount of α helices. However, the largest portion of this enzyme is composed of other domains. Besides, the amino acid sequence does not appear to be highly conserved, as there are at least as many significantly low-conservation segments as there are high-conservation ones (Figure 4). To determine what other domains apart from multiple β sheet and only few α helix regions form the protein, its sequence was inspected for the presence of potential coiled-coil motifs, which proved not to be the case.

Next, the databases Pfam and PDB were browsed in parallel to examine the topology of Granzyme with a holistic approach. On the one hand, the results clarified that Granzyme B is composed of one Trypsin motif (length: 219 amino acids) as well as several small disordered fragments (Figure 5). On the other hand, it was shown that the protein arises from the combination of two homologous subunits hold together via

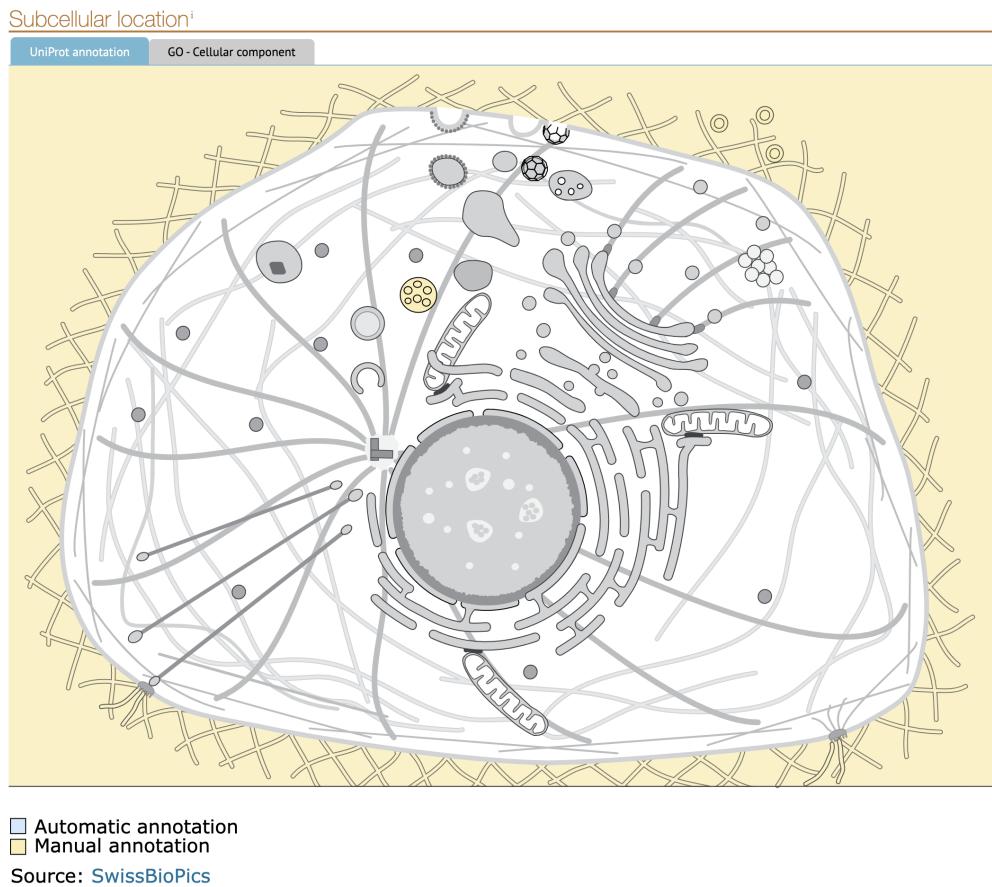


Figure 3: Subcellular location of Granzyme B. The yellow regions correspond to the extracellular matrix and the cytosolic granules of T cells, respectively. The enzyme is shipped to the cell membrane with a vesicle and then transferred to the target cell through an immunological synapsis.

intermolecular interactions (Figure 6).

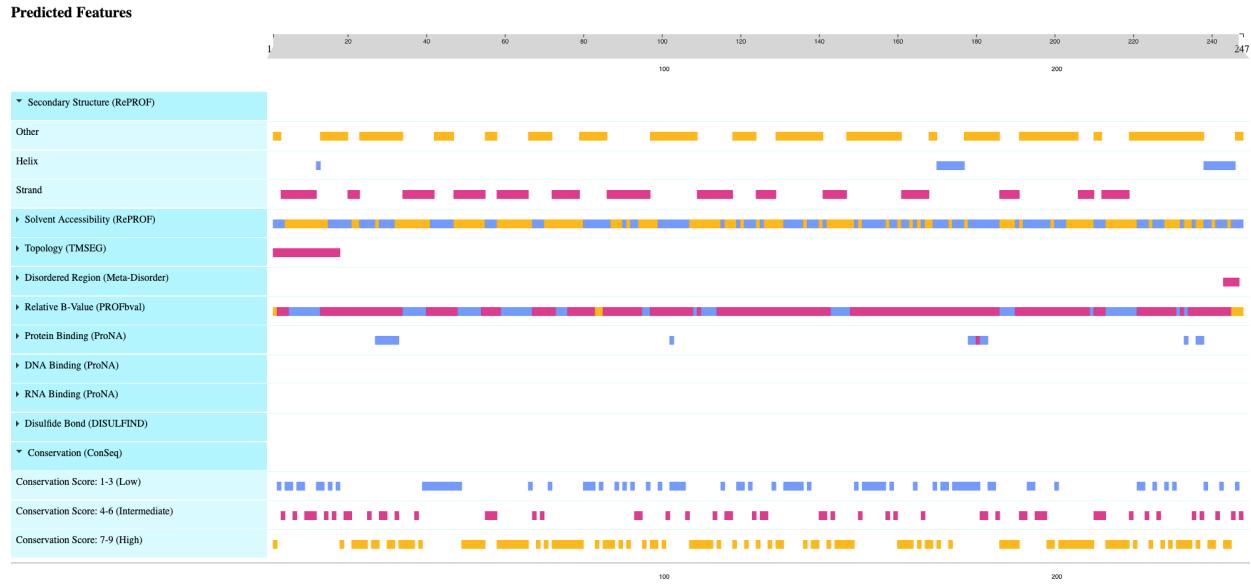


Figure 4: Predicted features for the entire length of human Granzyme B (247 amino acids). Predictions of secondary structure, conservation, protein binding and other properties were returned by PredictProtein. The colour scales are described in detail within the each category.

Source	Domain	Start	End
sig_p	n/a	1	18
low_complexity	n/a	5	13
Pfam	Trypsin	21	240
disorder	n/a	85	88
disorder	n/a	90	93
disorder	n/a	122	132
disorder	n/a	147	148
disorder	n/a	158	162
disorder	n/a	164	168

Figure 5: Tabular overview on elementary domains forming Granzyme B. The Trypsin motif spans the largest region (length: 219 amino acids), whereas disordered fragments of 2-10 amino acids are evenly distributed over the sequence

0.0.10 4.3. comparison with Granzyme H Secondary structure alignments help visualise the regions of high and low similarity between two enzymes, which catalise a common reaction or share the analogous biochemical properties. In this way, it is possible to draw conclusions on the location of their binding sites and their mode of action. As reported in Figure 7, the human protease Granzyme H resulted as a high similarity hit of Granzyme B through BLAST and PSI-BLAST across the PDB database, therefore the structures of the two proteins were aligned and compared in terms of sequence and structure conservation (Figures 8 and 9). Despite a relatively high sequence variability at some particular sites, the secondary structure is firmly conserved.

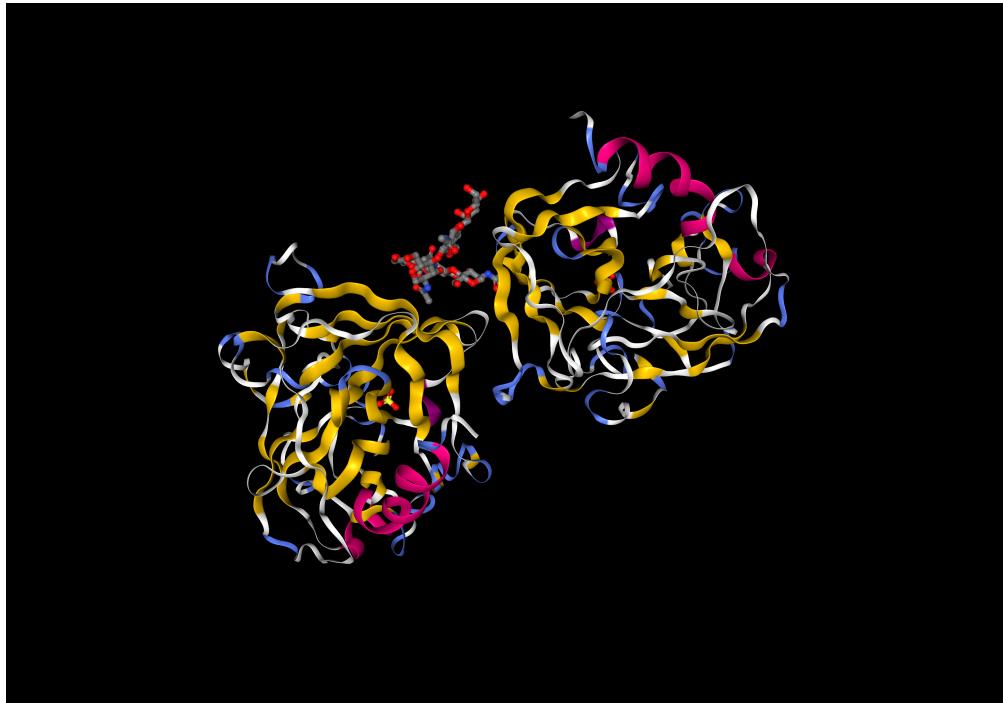


Figure 6: Chrystallographic reconstruction of Granzyme B. The two subunits are coloured by secondary structure: yellow stands for strands, magenta for helices, blue for disordered regions and violet for other.

No 1: Query=1fq3A Sbjct=3tk9A Z-score=39.7

[back to top](#)

DSSP LLLLLLLLLLLEEEELLEEEELLEEEELLEEEELHHHLLLEEEEELLLLLL
Query IIGGHEAKPHSRPYMAYLMIWDQKSILRKGGLIQDFVLTAACWGSINVTLGAHNK 60
ident ||||||| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct IIGGHEAKPHSRPYMAFVQLQEKSRKGGLILVRKDVFVLTAACQGSINVTLGAHNK 60
DSSP LLLLLEELLLLLLLEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLL
DSSP LLLLLLEEEEEEEELLLLLLLEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLL
Query QEERTQQFIPVKRPIPHAYNPKNFSNDIMLQLERKAKRTRAVPQLRLPSNKQAQVKPGQ 120
ident ||||| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct QEERTQQFIPVKRPIPHAYNPKNFSNNIMLQLERKAKWTTAVRPLRLPSSKAQVKPGQ 120
DSSP LLLLLEEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLL
DSSP EEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLL
Query TCSVAGWGQTAPLGhSHTLQEVKMTVQEDRKCESDLRHYYDSTIELCVGDPEIKTSFK 180
ident ||||||| | | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct LCSVAGWGYZMSMTI-ATTLQEVLLTVQKDQCERLFHGNYSRATELCVGDPKKTQGFK 179
DSSP EEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLL
DSSP LLLLLLLLLLLEEEELLLLLLLEEEELHHHHHHHHHHHHHHHHHHHH
Query GDGGPLVCKNVAQGIVSYGRNNGMPRACTKVSSFVHWIKKTMKRY 227
ident ||||||| | | | | | | | | | | | | | | | | | | | | | | | | | | | |
Sbjct GDGGPLVCKDVAQGILSGYNKKGTPPGVYIKVSHFLPIKRTMKRY 226
DSSP LLLLLEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLLLEEEEEEELLLL

Figure 7: Sequence alignment of human Granzymes B and H. The amino acid sequences of Granzymes B and H correspond to the upper and the lower strands, respectively. The letters above and below the alignment indicate whether the amino acid most likely lies within a helix (H), strand (E) or other region (L). Amino acids from position 121 on seem to match less ideally than those upstream.

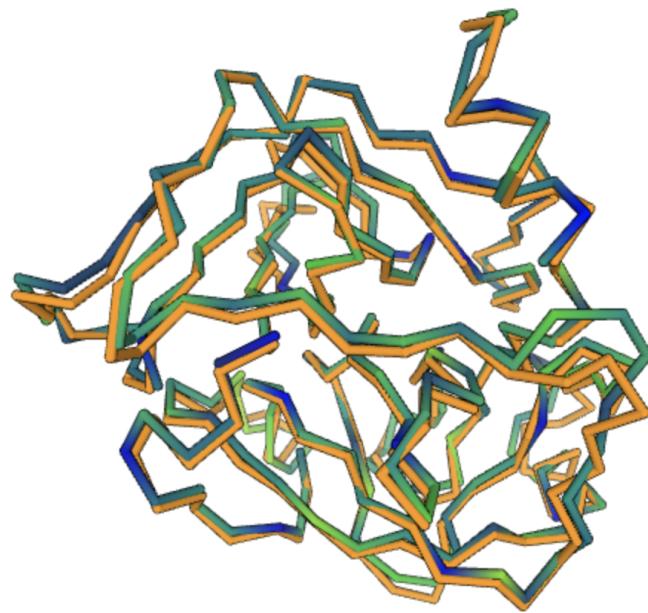


Figure 8: Pairwise alignment by sequence conservation. Granzyme B is represented with the orange backbone and Granzyme H with the green one. The multiple blue sites reflect regions of low similarity between the sequences.

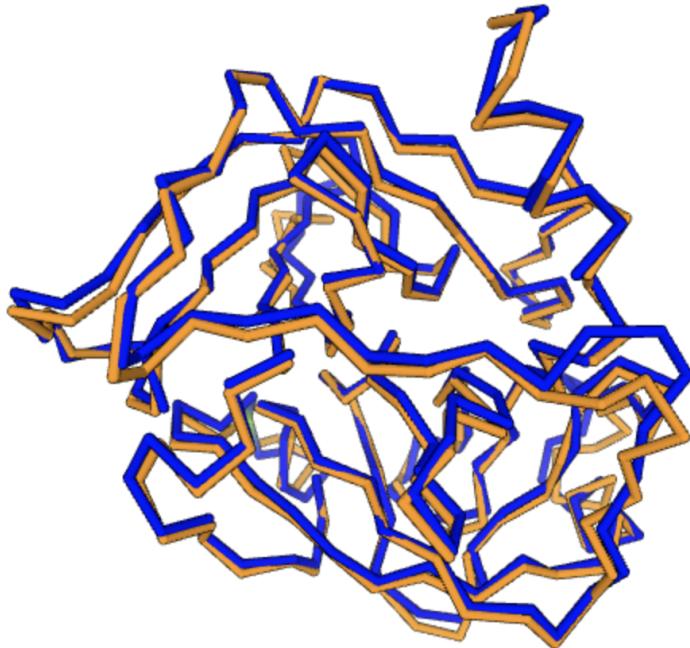


Figure 9: Pairwise alignment by structure conservation. Granzyme B is represented with the orange backbone and Granzyme H with the blue one. The few green sites reflect regions of low similarity between the structures.

4.4. results of MSA The sequences of six orthologs and one variant of human Granzyme B were selected so that the MSA could account for various degrees of similarity with respect to human Granzyme B. Figure 10 illustrates that, despite a relatively high variability, structural and biochemical properties are preserved across the orthologs. Moreover, the initial segments of most sequences (from position 3 to 15) exhibit a common repeated pattern which is cleaved when the precursor protein becomes mature.

The obtained MSA was then employed to reconstruct the phylogenetic tree of the protein orthologs, which well reflects the evolutionary relationships among the selected mammalian species (Figure 11). Additionally, the sequence reconstruction was validated by comparison with the analogous clustalw dendrogram.

CLUSTAL O(1.2.4) multiple sequence alignment

```

NP_612526.2          --MKLLLLLSFSLAPKTEAGEIIGGHEAKPHSRPYMAYLQIMDE  

NP_038570.1          --MKILLLLTLSLARTKAGEIIGGHEVPHSRPYMALLIKDQO  

NP_001254677.1      --MQPILLLLLFLLPRADGEIGGHEAKPHSRPYMAYLQIMDE  

NP_001292863.1      --MQPILLLLLFLLPRADGEIGGHEAKPHSRPYMAYLWIMDD  

NP_004122.2          --MQPILLLLLFLLPRADGEIGGHEAKPHSRPYMAYLWIMDD  

NP_001332940.1      --MQPILLLLLFLLPRAEGEIGGHEAKPHSRPYMAYLQIQDQO  

NP_001137182.1      --MQPILLLLLFLLPSGTENEIGGHEAKPHSRPYMAYLQVFLE  

NP_001075350.1      ;*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:  
  

NP_612526.2          FVLTAHCGSGSKINVTLGAHNIKQEMQOIIPVKIPHPAYNSK  

NP_038570.1          FVLTAHCGECGSGSInvTLGAHNIKQEMQOIIPVKIPHPDYNK  

NP_001254677.1      FVLTAHCGWGSInvTLGAHNIKQEMQOIIPVKIPHPDYNK  

NP_001292863.1      FVLTAHCGWGSInvTLGAHNIKQEMQOIIPVKIPHPDYNK  

NP_004122.2          FVLTAHCGWGSInvTLGAHNIKQEMQOIIPVKIPHPDYNK  

NP_001332940.1      FVLTAHCGWGSInvTLGAHNIKQEMQOIIPVKIPHPDYNK  

NP_001137182.1      FVLTAHCGWGSInvTLGAHNIKQEMQOIIPVKIPHPDYNK  

NP_001075350.1      FVLTAHCGWGSInvTLGAHNIKQEMQOIIPVKIPHPDYNK  

;*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:  
  

NP_612526.2          AKRSPAVPLNLRLRNNVKVPGDCVCYVAGWGKLGPGNYSDLQEVE  

NP_038570.1          AKRTRAVPQLPLRNNVKVPGDCVCYVAGWGKLGPGNYSNTLQEV  

NP_001254677.1      AKRTRAVPQLPLRNNSKAVQPQCVSGAVGWGTRPTYSHTLQAVE  

NP_001292863.1      AKRTRAVPQLPLRNNSKAVQPQCVSGAVGWGTRPTYSHTLQAVE  

NP_004122.2          AKRTRAVPQLPLRNNSKAVQPQCTVSAVGWGOTAPLGHSTLQEV  

NP_001332940.1      AKRTRAVPQLPLRNNSKAVQPQCTVSAVGWGOTAPLGHSTLQEV  

NP_001137182.1      AKLTRAVTTLSRLRAKRQAVPIQCVRGWGQTPALQHSTLQEV  

NP_001075350.1      AKLTRAVTTLSRLRAKRQAVPIQCVRGWGQTPALQHSTLQEV  

;*****:*****:*****:*****:*****:*****:*****:*****:*****:*****:  
  

NP_612526.2          KNYFDKANEICAGDPKIKRASFRDSGPGLVCKKVAAGIVSYGQNDE  

NP_038570.1          KNYRNTNOICGDPKIKRASFRDSGPGLVCKKVAAGIVSYGYCDE  

NP_001254677.1      RNYDSTEVELCGDVEIPKKASFRDSGPGLVCKKVAAGIVSYGHPGINCE  

NP_001292863.1      GHYYDSTEVELCGDVEIPKKASFRDSGPGLVCKKVAAGIVSYGQNGE  

NP_004122.2          RHYDSTEVELCGDVEIPKKASFRDSGPGLVCKKVAAGIVSYGRNNE  

NP_001332940.1      RHYYDSTEVELCGDVEIPKKASFRDSGPGLVCKKVAAGIVSYGRNNE  

NP_001137182.1      PNYYNSTTQLCGDPKKKKQATFGDSGPGLVCKKVAAGIVSYGRNNE  

NP_001075350.1      PNYYNSTTQLCGDPKKKKQATFGDSGPGLVCKKVAAGIVSYGRNNE  

;*****:*****:*****:*****:*****:*****:*****:*****:*****:  
  

NP_612526.2          SWIKKTMMKKS-----248  

NP_038570.1          SWIKKTMMKKS-----247  

NP_001254677.1      HWIKKTMRPGTMAYTCNPTRLGGRWIRSRV270  

NP_001292863.1      RWIKKTMRKRH-----247  

NP_004122.2          HWIKKTMMKRY-----247  

NP_001332940.1      HWIKKTMMKRY-----235  

NP_001137182.1      PWIKKTMMSL-----248  

NP_001075350.1      PWIKKTMMKL-----247  

;*****:*****:  


```

Figure 10: MSA among several orthologs of Granzyme B. Colours refer to the clustalx scale and provide a measure of structure conservation.

5. Conclusions

Generally speaking, this analysis draws to the following points:

1. The Tripsin motif could be responsible for the proteolytic activity of Granzyme B;
 2. Proteins from the Granzyme family, such as Granzyme B and H, exhibit a high structure conservation despite some differences at the sequence level;
 3. Granzyme B is very recurrent in nature as an essential element of the immune system of several animals;
 4. The sequence of the precursor Granzyme B contains an initial repetitive segment which is maintained across several animals.

The Tripsin motif was first observed in the serine protease with the same name.²⁶ This enzyme is commonly found in the digestive system of multiple vertebrate organisms, where it catalyses protein hydrolysis. Because

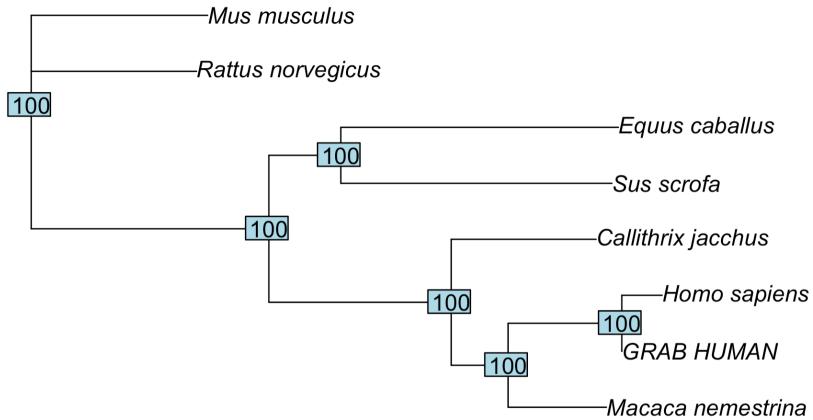


Figure 11: Sequence reconstruction of the Granzyme family. The tree was generated with the upgma algorithm and rooted with *M. musculus* and *R. norvegicus* as the outgroups. Bootstrapping values are reported on top of the corresponding branches. The element GRAB HUMAN refers to human Granzyme B, whereas *H. sapiens* to isoform 2 of the same enzyme.

the Tripsin motif is also present within Granzyme B (Figure 5), this element could be responsible for the proteolytic activity at the protein binding site of the enzyme.

Interestingly, from Figure 4 it is possible to infer that protein binding site of Granzyme B does not fully coincide with the highly conserved regions of the sequence. This feature might lead to a relative substrate unspecificity, which could bring benefits in the fight against pathogens with different proteomes.

Moreover, the sequence variability of the binding site, if present in other members of the Granzyme family, could explain why different Granzymes bind to and catalyse the hydrolysis of different proteins. What was unraveled, however, is that changes in the amino acid sequence of these enzymes do not remarkably affect their secondary structures (Figures 8 and 9). This aspect should be subject of deeper analysis in the future.

The Granzyme family plays an essential role in the adaptive immune response against viral and bacterial intracellular pathogens, therefore its presence and conservation among several animals, such as those included in Figure 11, is not surprising. Additionally, the initial DNA fragment which defines a precursor protein of Granzyme B is conserved across most animals. This suggests that such organisms might coordinate the transcription as well as post-translational modifications of this proteins through analogous mechanisms.

Taken all together, this report shows the incredibly large amount of information on a protein that a structural analysis *in silico* can convey. In the future, this type of investigation will likely cover an increasingly important position in the discovery of new drugs, the mining of usable natural compounds as well as the response to the climate crisis.

6. References

1. Dayhoff, M., Schwartz, R. & Orcutt, B. 22 a model of evolutionary change in proteins. *Atlas of protein sequence and structure* **5**, 345–352 (1978).
2. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915–10919 (1992).

3. Chou, P. Y. & Fasman, G. D. Prediction of protein conformation. *Biochemistry* **13**, 222–245 (1974).
4. Krähenbühl, O. *et al.* Characterization of granzymes a and b isolated from granules of cloned human cytotoxic t lymphocytes. *The Journal of Immunology* **141**, 3471–3477 (1988).
5. Hameed, A., Lowrey, D., Lichtenheld, M. & Podack, E. Characterization of three serine esterases isolated from human il-2 activated killer cells. *The Journal of Immunology* **141**, 3142–3147 (1988).
6. Poe, M. *et al.* Human cytotoxic lymphocyte granzyme b. Its purification from granules and the characterization of substrate and inhibitor specificity. *Journal of Biological Chemistry* **266**, 98–103 (1991).
7. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).
8. Altschul, S. F. *et al.* Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402 (1997).
9. Apweiler, R. *et al.* UniProt: The universal protein knowledgebase. *Nucleic acids research* **32**, D115–D119 (2004).
10. Rost, B., Yachdav, G. & Liu, J. The predictprotein server. *Nucleic acids research* **32**, W321–W326 (2004).
11. Gasteiger, E. *et al.* Protein identification and analysis tools on the expasy server. *The proteomics protocols handbook* 571–607 (2005).
12. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* 1162–1164 (1991).
13. Bateman, A. *et al.* The pfam protein families database. *Nucleic acids research* **32**, D138–D141 (2004).
14. Sussman, J. L. *et al.* Protein data bank (pdb): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography* **54**, 1078–1084 (1998).
15. Estébanez-Perpiñá, E. *et al.* Crystal structure of the caspase activator human granzyme b, a proteinase highly specific for an asp-p1 residue. (2000).
16. Wang, L. *et al.* Structural insights into the substrate specificity of human granzyme h: The functional roles of a novel rkr motif. *The Journal of Immunology* **188**, 765–773 (2012).
18. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL w: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673–4680 (1994).
19. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology* **7**, 539 (2011).
20. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
21. Schliep *et al.* Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution* **8**, 1212–1220 (2017).
22. Morgan, M. *BiocManager: Access the bioconductor project package repository*. (2021).
23. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. in *Structural approaches to sequence evolution: Molecules, networks, populations* (eds. Bastolla, U., Porto, M., Roman, H. E. & Vendruscolo, M.) 207–232 (Springer Verlag, 2007).
24. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2021).

25. RStudio Team. *RStudio: Integrated development environment for r*. (RStudio, PBC, 2020).
26. Koshikawa, N., Yasumitsu, H., Nagashima, Y., Umeda, M. & Miyazaki, K. Identification of one-and two-chain forms of trypsinogen 1 produced by a human gastric adenocarcinoma cell line. *Biochemical Journal* **303**, 187–190 (1994).